e considerando i soggetti di ogni grappolo come gli elementi del campione. Il campione **a più stadi** è ottenuto attraverso combinazioni di questi metodi.

Nel Capitolo 3 verranno presentate le statistiche che si impiegano per descrivere i campioni e i corrispondenti parametri per descrivere le popolazioni. Il tema centrale del capitolo sarà quindi la *statistica descrittiva*.

Problemi

Concetti di base

2.1 Spiega le differenze fra

a. Variabili discrete e continue.
b. Variabili categoriali e quantitative.
c. Variabili ordinali e nominali.

Perché queste distinzioni sono importanti per l'analisi statistica?

2.2 Classifica ciascuna variabile come categoriale o quantitativa.

a. Numero di animali posseduti dalla famiglia.
b. Nazionalità.
c. Scelta dell'auto (nazionale o d'importazione).
d. Distanza (in km) per raggiungere il posto di lavoro.
e. Tipo di dieta (vegetariana o non vegetariana).
f. Tempo trascorso navigando su Internet nell'ultimo mese.
g. Possesso di un personal computer (sì, no).
h. Numero di persone conosciute malate di AIDS (0, 1, 2, 3, 4 o più).
i. Tipologia delle unioni matrimoniali in una società (monogamia, poligamia, poliandria).

2.3 Quale scala di misura (nominale, ordinale o di intervalli) è più appropriata per misurare

a. Opinione sulla legalizzazione del consumo della marijuana (favorevole, neutrale, contrario).
b. Sesso (femmina, maschio).
c. Numero di figli in una famiglia (0, 1, 2, . . .).
d. Affiliazione partitica (Democratico, Repubblicano, Indipendente).
e. Affiliazione religiosa (cattolico, ebreo, protestante, musulmano, altro).
f. Ideologia politica (molto liberale, moderatamente liberale, moderato, moderatamente conservatore, conservatore).
g. Anni di scolarizzazione (0, 1, 2, 3, . . .).
h. Titolo di studio (nessuno, scuola superiore, laurea di primo livello, laurea di secondo livello, master, dottorato).
i. Ambito degli studi universitari (umanistico, scientifico, sociale, tecnico, sanitario).
j. Punteggio ottenuto in un test (intervallo 0-100).
k. Status occupazionale (occupato a tempo pieno, occupato part-time, disoccupato).

2.4 Quale scala di misura è più appropriata per misurare

a. Professione (idraulico, insegnante, impiegato, . . .).
b. Status della professione (esecutivo, di concetto).
c. Status sociale (basso, medio, alto).
d. Tasso di omicidi (numero di omicidi ogni 1000 abitanti).
e. Popolazione residente in una regione (numero di residenti).
f. Tasso di crescita di una popolazione (in percentuale).
g. Classificazione di un comune (rurale, piccola città, grande città, metropoli).
h. Reddito annuo (in migliaia di euro).
i. Atteggiamento verso le azioni finalizzate a combattere le discriminazioni (favorevole, neutrale, non favorevole).
j. Numero di partner con cui si sono avute relazioni sessuali.

2.5 Qual è la scala di misura più appropriata per la variabile "livello di istruzione" misurata come

a. Numero di anni (0, 1, 2, 3, . . .).
b. Titolo conseguito (elementare, medio inferiore, medio superiore, laurea).
c. Tipologia di scuola frequentata (pubblica, privata).

2.6 Fai alcuni esempi per una variabile che è (*a*) categoriale, (*b*) quantitativa, (*c*) ordinale, (*d*) nominale, (*e*) discreta, (*f*) continua, (*g*) quantitativa discreta.

2.7 In un'indagine condotta nel giugno 2006 da YouGov per il giornale britannico *The Daily Telegraph* sono state sottoposte a un campione di 1962 adulti inglesi alcune domande per valutare la loro opinione sugli Stati Uniti. Uno dei quesiti era "Qual è il tuo giudizio su George W. Bush come leader?", le possibili risposte erano: grande leader, leader abbastanza grande, piuttosto scarso come leader, pessimo leader.

a. Questa variabile a quattro categorie è nominale o ordinale? Perché?
b. Si tratta di una variabile continua o discreta? Perché?
c. Per il 93% degli intervistati che ha risposto al quesito si sono osservate le seguenti percentuali nelle quattro categorie: 1% (grande leader), 16%, 37%, 46% (pessimo leader). Questi valori sono statistiche o parametri? Perché?

2.8 In un'indagine è stato chiesto ad alcuni soggetti di attribuire un giudizio di importanza a cinque diverse tematiche in grado di influire sulle loro intenzioni di voto per il senato USA. La scala utilizzata era: molto importante, importante, non importante. Le cinque tematiche erano: politica estera, disoccupazione, inflazione, corsa agli armamenti e diritti civili. Le valutazioni espresse possono essere considerate come cinque variabili: valutazione sulla politica estera, valutazione sulla disoccupazione e così via. Che scala di misurazione rappresentano queste variabili?

2.9 Quale delle seguenti variabili potrebbe essere teoricamente misurata su una scala continua? (*a*) Metodo di contraccezione impiegato, (*b*) durata della residenza in una certa città, (*c*) tempo di completamento di un test, d'intelligenza, (*d*) autoritarismo, (*e*) alienazione, (*f*) regione di residenza.

2.10 Quale delle seguenti variabili è continua se la sua misurazione avviene con la massima precisione possibile? (*a*) Età della madre, (*b*) numero di figli nella famiglia, (*c*) reddito del coniuge, (*d*) popolazione residente in una città, (*e*) latitudine e longitudine di un luogo, (*f*) distanza tra luogo di lavoro e luogo di residenza per un individuo, (*g*) numero di lingue straniere conosciute.

2.11 Una classe è costituita da 50 studenti. Impiegando la colonna delle prime due cifre nella tavola dei numeri casuali riportata nella Tabella 2.1 seleziona un campione casuale di tre studenti. Se gli studenti fossero stati numerati da 01 a 50 quali sarebbero i numeri dei tre studenti selezionati?

2.12 L'elenco telefonico di una città contiene 400 pagine ciascuna delle quali contiene 130 nomi per un totale di 52 000 nomi. Spiega come potresti selezionare un campione casuale semplice di 5 nomi. Utilizzando la seconda colonna della Tabella 2.1 oppure un software o un calcolatore seleziona 5 numeri casuali per identificare i soggetti del campione.

2.13 Spiega perché un esperimento oppure uno studio osservazionale sarebbero più appropriati per indagare sui seguenti temi.

a. Se le città con un elevato tasso di disoccupazione hanno anche un più elevato tasso di criminalità.
b. Se la Honda Accord ha un consumo inferiore della Toyota Camry.
c. Se gli studenti che hanno voti elevati al diploma di scuola media superiore tendono ad avere voti più alti negli studi universitari.
d. Se una speciale cartolina riportata nella copertina di un catalogo rende più probabile che i lettori ordinino dei prodotti reclamizzati nello stesso catalogo.

2.14 Uno studio è stato pianificato per verificare se il fumo passivo è responsabile di alte incidenze di tumori all'apparato respiratorio.

a. Un possibile studio prevede la selezione di un campione di bambini. A caso, la metà di loro viene fatta stare in un ambiente in cui vi sono fumatori, l'altra metà in un ambiente in cui non si trovano fumatori. Dopo 60 anni di osservazione viene stabilito quanti di loro hanno sviluppato dei tumori a carico dell'apparato respiratorio. Sarebbe questo uno studio sperimentale o osservazionale? Perché?
b. Per diverse ragioni, fra le quali la disponibilità di tempo e i codici etici, non è possibile condurre degli studi come quelli descritti nel punto (*a*). Descrivi un modo in cui sarebbe possibile studiare gli effetti del fumo passivo specificando se si tratta di uno studio sperimentale oppure osservazionale.

2.15 La Tabella 2.2 mostra i risultati delle elezioni presidenziali USA e gli stessi risultati così come erano stati previsti da diversi istituti demoscopici nei giorni immediatamente precedenti le elezioni. Le ampiezze campionarie considerate erano, generalmente, intorno alle 2000 unità. La percentuale per ogni sondaggio non somma a 100 perché certi elettori si dichiaravano indecisi o a favore di altri candidati diversi dai tre riportati nella tabella.

a. Quali fattori spiegano la variabilità dei risultati previsti dai vari istituti?
b. Identifica l'errore campionario per il risultato previsto dalla Gallup.

Tabella 2.2

Sondaggio	Voti previsti		
	Gore	Bush	Nader
Gallup	46	48	4
Harris	47	47	5
ABC	45	48	3
CBS	45	44	4
NBC	44	47	3
Pew Research	47	49	4
Voto effettivo	48.4	47.9	2.7

Fonte. www.ncpp.org/

2.16 LA BBC ha chiesto ai suoi telespettatori inglesi di chiamare telefonicamente il network per indicare la loro opera di poesia preferita. Degli oltre 7500 spettatori che hanno chiamato, la stragrande maggioranza ha indicato come opera preferita *If* di Rudyard Kipling. La BBC ha presentato la notizia dicendo che Kipling era l'autore chiaramente preferito.

a. Spiega perché quello della BBC è chiamato "campione volontario".
b. Volendo effettivamente determinare quella che era l'opera di poesia preferita dagli inglesi come avrebbe potuto agire la BBC?

assunti dalle osservazioni. I valori della variabile risposta sono rappresentati sull'asse y mentre quelli della variabile esplicativa sull'asse x.

- Per due variabili quantitative, la **correlazione** è una misura della forza del legame di associazione lineare fra le variabili. Essa assume valori fra -1 e $+1$ e indica se i valori della variabile risposta tendono ad aumentare (correlazione positiva) o a diminuire (correlazione negativa) quando aumentano i valori della variabile esplicativa.
- L'**analisi della regressione** ci permette di ottenere l'equazione di una retta per prevedere i valori di una variabile risposta sulla base dei valori assunti da una variabile esplicativa. Nel Capitolo 9 studieremo nel dettaglio la correlazione e la regressione.

Problemi

Concetti di base

3.1 La Tabella 3.10 mostra i valori (espressi in milioni) della popolazione nata all'estero residente negli USA nel 2004 ripartita secondo il luogo di nascita.

a. Costruisci la distribuzione delle frequenze relative.
b. Rappresenta la distribuzione attraverso un grafico a barre.
c. La variabile "Luogo di nascita" è quantitativa o categoriale?
d. Impiega qualunque misura ritieni sia utile per sintetizzare i dati: media, mediana o moda.

Tabella 3.10

Luogo di nascita	Numerosità
Europa	4.7
Asia	8.7
Caraibi	3.3
America Centrale	12.9
Sud America	2.1
Altro	2.6
Totale	**34.3**

3.2 Secondo quanto riportato su www.adherents.com nel 2006 il numero di credenti delle prime cinque religioni del mondo era 2.1 miliardi di cristiani, 1.3 miliardi di islamici, 0.9 miliardi di induisti, 0.4 miliardi di confuciani, 0.4 miliardi di buddisti.

a. Costruisci la distribuzione delle frequenze relative.
b. Rappresenta la distribuzione attraverso un grafico a barre.
c. Puoi individuare la media, la mediana o la moda per questi dati? Se sì, determina tali grandezze e commenta.

3.3 Un insegnante presenta alla sua classe il diagramma ramo-e-foglie della distribuzione dei voti ottenuti dagli studenti in un semestre:

```
6 | 5 8 8
7 | 0 1 1 3 6 7 7 9
8 | 1 2 2 3 3 3 4 6 7 7 8 9
9 | 0 1 1 2 3 4 4 5 8
```

a. Determina il totale degli studenti e i punteggi massimo e minimo.
b. Rappresenta la distribuzione attraverso un istogramma per quattro intervalli di voti.

3.4 Secondo la *2005 American Community Survey*, nel 2005 negli USA si sono registrate 30.1 milioni di famiglie con un solo componente, 37.0 milioni con due componenti, 17.8 milioni con tre componenti, 15.3 milioni con quattro componenti e 10.9 milioni con 5 o più componenti.

a. Costruisci la distribuzione delle frequenze relative.
b. Rappresenta la distribuzione attraverso un istogramma. Che forma ha?
c. Determina e interpreta i valori della mediana e della moda della variabile "ampiezza delle famiglie".

3.5 Copia il file di dati "2005 statewide crime" dal sito web del testo (www.stat.ufl.edu/~aa/social/data.html). Impiegando un software, fai riferimento alla variabile tasso di omicidi (murder rate) per 100 000 abitanti (in questo problema non considerare l'osservazione riferita al District of Columbia – D.C.).

a. Costruisci la distribuzione delle frequenze relative.
b. Costruisci l'istogramma della distribuzione. Che forma ha?
c. Costruisci il diagramma ramo-e-foglie. Come appare in confronto con l'istogramma ottenuto in (*b*)?

3.6 La OECD (*Organization for Economic Cooperation and Development*) riunisce i paesi più avanzati e industrializzati del mondo che accettano i principi della democrazia rappresentativa e dell'economia basata sul libero mercato. La Tabella 3.11 mostra i dati raccolti dalle Nazioni Unite (UN) per i paesi OECD con riferimento a diverse variabili: Prodotto Interno Lordo (PIL) pro-capite in dollari USA (in inglese, GDP = *Gross Domestic Product*), percentuale di disoccupati, misura della disuguaglianza sociale basata sul confronto fra la ricchezza posseduta dal 10% più ricco e dal 10% più povero della popolazione, spesa pubblica per la sanità (in percentuale del PIL), numero di medici per 100 000 persone, emissioni di anidride carbonica (in tonnellate pro capite), percentuale di seggi parlamentari occupati da donne, donne occupate ogni 100 lavoratori maschi.

Tabella 3.11 Dati UN per le nazioni OECD. Disponibile come file "OECD data" sul sito web del testo.

Nazione	PIL	Disoc.	Disuguag.	Salute	Medici	CO_2	Donne parl.	Fem. econ.
Australia	30,331	5.1	12.5	6.4	247	18	28.3	79
Austria	32,276	5.8	6.9	5.1	338	8.6	32.2	75
Belgio	31,096	8.4	8.2	6.3	449	8.3	35.7	72
Canada	31,263	6.8	9.4	6.9	214	17.9	24.3	83
Danimarca	31,914	4.9	8.1	7.5	293	10.1	36.9	84
Finlandia	29,951	8.6	5.6	5.7	316	13	37.5	86
Francia	29,300	10.0	9.1	7.7	337	6.2	13.9	79
Germania	28,303	9.3	6.9	8.7	337	9.8	30.5	76
Grecia	22,205	10.6	10.2	5.1	438	8.7	13	66
Islanda	33,051	2.5	..	8.8	362	7.6	33.3	87
Irlanda	38,827	4.3	9.4	5.8	279	10.3	14.2	72
Italia	28,180	7.7	11.6	6.3	420	7.7	16.1	61
Giappone	29,251	4.4	4.5	6.4	198	9.7	10.7	65
Lussemburgo	69,961	4.6	..	6.2	266	22	23.3	68
Paesi Bassi	31,789	6.2	9.2	6.1	315	8.7	34.2	76
Nuova Zelanda	23,413	3.6	12.5	6.3	237	8.8	32.2	81
Norvegia	38,454	4.6	6.1	8.6	313	9.9	37.9	87
Portogallo	19,629	7.5	15	6.7	342	5.6	21.3	79
Spagna	25,047	9.1	10.3	5.5	330	7.3	30.5	65
Svezia	29,541	5.6	6.2	8	328	5.9	45.3	87
Svizzera	33,040	4.1	9	6.7	361	5.6	24.8	79
Regno Unito	30,821	4.8	13.8	6.9	230	9.4	18.5	79
USA	39,676	5.1	15.9	6.8	256	19.8	15	81

Fonte: hdr.undp.org/statistics/data
Disoc. = % Disoccupazione, Disuguag. = Misura di disuguaglianza, Donne parl. = % di seggi parlamentari occupati da donne, Fem. econ. = Attività economica femminile (% degli occupati maschi).

a. Costruisci il diagramma ramo-e-foglie per i valori del PIL arrotondando gli stessi e esprimendoli in migliaia di dollari (ad esempio, sostituendo \$ 19 629 con 20).
b. Costruisci l'istogramma corrispondente al diagramma ramo-e-foglie ottenuto in (*a*).
c. Identifica l'outlier in ciascuna distribuzione.

3.7 Di recente, il numero di aborti per 1000 donne in età dai 15 ai 41 anni per gli stati della costa del Pacifico USA è stato: Washington, 26; Oregon, 17; California, 236; Alaska, 2; Hawaii, 6 (*Statistical Abstract of the United States, 2006*).

a. Calcola la media.
b. Calcola la mediana. Perché è così diversa dalla media?

3.8 Il fenomeno del riscaldamento globale viene in gran parte attribuito alle attività umane che provocano emissioni di anidride carbonica e di altri gas "effetto serra". Lo *Human Development Report 2005*, pubblicato nell'ambito del Programma di Sviluppo delle Nazioni Unite, riporta le emissioni (in tonnellate pro capite) registrate nel 2002 per le otto nazioni più popolate del pianeta: Bangladesh 0.3, Brasile 1.8, Cina 2.3, India 1.2, Indonesia 1.4, Pakistan 0.7, Russia 9.9, USA 20.1.

a. Calcola la media e la mediana di questi otto valori.
b. C'è qualche osservazione outlier? Discuti l'influenza esercitata dall'outlier sulla media e sulla mediana.

3.9 In un'indagine della organizzazione Roper è stato chiesto a un campione di intervistati quanto ritenessero avanzate le norme sulla protezione ambientale. Le possibili risposte "per niente avanzate", "adeguate" e "troppo avanzate" sono state scelte, rispettivamente, dalle seguenti percentuali di rispondenti 51%, 33%, 16%.

a. Qual è la moda delle risposte?
b. Possiamo calcolare una media o una mediana per questi dati? Se sì, fallo; altrimenti, spiega perché no.

3.10 Un ricercatore di un centro per il trattamento dell'alcolismo, al fine di studiare la lunghezza dei ricoveri nel centro dei pazienti non recidivi, seleziona casualmente dieci individui che sono stati ricoverati nei due anni precedenti. Le lunghezze delle loro degenze, espresse in giorni, sono risultate 11, 6, 20, 9, 13, 4, 39, 13, 44 e 7.

a. Costruisci il diagramma ramo-e-foglie.
b. Calcola media e deviazione standard e interpreta i risultati.
c. In uno studio simile, condotto 25 anni fa, la lunghezza delle degenze per un campione di 10 individui era risultata pari a 32, 18, 55, 17, 24, 31, 20, 40, 24, 15. Confronta questi risultati con quelli ottenuti nel nuovo studio utilizzando (i) un diagramma ramo-e-foglie "schiena contro schiena", (ii) la media, (iii) la deviazione standard. Interpreta le differenze che rilevi.
d. Nel nuovo studio è stata aggiunta al campione un'ulteriore osservazione riferita a una persona ricoverata da 40 giorni e ancora degente. La lunghezza della degenza di questo paziente è come minimo pari a 40 ma il suo effettivo valore è incognito. Puoi calcolare la media o la mediana per il campione di 11 osservazioni che include l'osservazione parziale? Spiega la tua risposta (un'osservazione come questa viene definita *censurata* per indicare che è un parziale di un incognito valore).

3.11 Accedi ai dati GSS all'indirizzo web http://sda.berkeley.edu/GSS. Digita TVHOURS per la variabile da cercare e year(2006) nella casella filtro; otterrai i dati delle ore giornaliere per persona passate a guardare la TV negli USA nel 2006.

a. Costruisci la distribuzione delle frequenze relative per i valori 0, 1, 2, 3, 4, 5, 6, 7 o più.
b. Descrivi la forma della distribuzione?
c. Spiega perché la mediana è 2?
d. La media è maggiore di 2. Perché?

3.12 La Tabella 3.12 mostra per l'anno 2003 il tasso di attività femminile (numero di donne che lavorano ogni 100 lavoratori di sesso maschile) per le nazioni dell'Europa Occidentale. Considerando anche i dati riferiti al Sud America riportati nella Tabella 3.4, costruisci il diagramma ramo-e-foglie "schiena contro schiena." Interpreta il risultato.

3.13 Secondo Statistics Canada la distribuzione del reddito annuo delle famiglie canadesi nel 2000 aveva una mediana di $ 46 752 e una media pari a $ 71 600. Come ritieni che sia la forma della distribuzione? Perché?

3.14 Nell'ambito della General Social Survey del 2006 è stata posta la seguente domanda "Quante volte hai avuto rapporti sessuali negli ultimi 12 mesi?" Nella Tabella 3.13 è riportata la distribuzione delle risposte date da 2333 intervistati.

Tabella 3.13

Frequenza rapporti sessuali	Frequenza
Nessuno	595
Uno o due	205
Circa uno al mese	265
2 o 3 volte al mese	361
Circa una volta alla settimana	343
2 o 3 volte alla settimana	430
Più di 3 volte alla settimana	134

a. Riporta la mediana e la moda e interpreta i risultati.
b. Considera la scala come se fosse quantitativa assegnando i punteggi 0, 0.1, 1.0, 2.5, 4.3, 10.8 e 17 alle diverse categorie per approssimare la frequenza mensile dei rapporti sessuali. Trova la media campionaria e interpreta i risultati.

3.15 Nel 2004, nell'ambito della GSS è stato chiesto ai rispondenti "Quanto spesso leggi il giornale?" Le possibili risposte erano (ogni giorno, alcune volte alla settimana, una volta alla settimana, meno di una volta alla settimana, mai). I conteggi, nelle diverse categorie sono risultati 358, 222, 134, 121, 71.

a. Identifica la moda e la mediana.
b. Sia $y =$ *numero* di volte che hai letto il giornale in una settimana. Assegnando alle categorie di risposta i valori 7, 3, 1, 0.5, 0 trova $\bar{y}$. Come è questo valore rispetto alla media di 4.4 registrata nella GSS del 1994.

Tabella 3.12 Dati UN per le nazioni OECD. Disponibile come file "OECD data" sul sito web del testo.

Nazione	Attività econ. femm.	Nazione	Attività econ. femm.	Nazione	Attività econ. femm.
Austria	66	Germania	71	Norvegia	86
Belgio	67	Grecia	60	Portogallo	72
Cipro	63	Irlanda	54	Spagna	58
Danimarca	85	Italia	60	Svezia	90
Finlandia	87	Lussemburgo	58	Regno Unito	76
Francia	78	Paesi Bassi	68		

Fonte. Human Development Report, 2005, United Nations Development Programme.

3.16 Secondo la *2005 American Community Survey* condotta dallo US Bureau of the Census, i guadagni mediani negli ultimi 12 mesi sono stati pari a $ 32 168 per le lavoratrici e $ 41 965 per i lavoratori. I guadagni medi sono stati pari a $ 39 890 per le femmine e $ 56 724 per i maschi.

a. Questi dati suggeriscono che le distribuzioni dei guadagni per ciascun sesso siano simmetriche o asimmetriche (asimmetria positiva o negativa)? Spiega la tua risposta.
b. I dati riportati sono riferiti a 73.8 milioni di femmine e a 83.4 milioni di maschi. Calcola il guadagno medio complessivo.

3.17 Nel 2003 negli Stati Uniti, il reddito familiare mediano era pari a $ 55 800 per le famiglie bianche, $ 34 400 per le famiglie nere e $ 34 300 per le famiglie ispaniche (*Statistical Abstract of the United States, 2006*).

a. Identifica la variabile risposta e la variabile esplicativa di questa analisi.
b. Combinando i dati dei tre gruppi disponiamo di informazioni sufficienti per calcolare la mediana? Perché?
c. Se i dati riportati fossero stati delle medie, cos'altro ti sarebbe servito per calcolare la media generale?

3.18 Nell'ambito della GSS è stato chiesto agli intervistati "Negli ultimi 12 mesi quanta gente hai conosciuto personalmente che è stata vittima diretta o indiretta di un omicidio?" La Tabella 3.14 mostra l'output di un'analisi condotta sulle risposte.

a. La distribuzione è campanulare, asimmetrica positiva o asimmetrica negativa?
b. Possiamo applicare la regola empirica a questa distribuzione? Perché?
c. Riporta il valore mediano. Se 500 osservazioni si spostassero dallo 0 al 6, come cambierebbe la mediana? Tutto ciò, che proprietà illustra della mediana?

3.19 Nell'ottobre del 2006 un articolo sui "Salari minimi" apparso su wikipedia.org riportava (in dollari USA) il salario minimo orario in cinque nazioni: $ 10.00 in Australia, $ 10.25 in Nuova Zelanda, $ 10.46 in Francia, $ 10.01 nel Regno Unito, $ 5.15 negli USA. Trova la mediana, la media, il campo di variazione e la deviazione standard (*a*) escludendo gli USA, (*b*) per tutte e cinque le osservazioni. Usa questi dati per spiegare gli effetti esercitati degli outlier su queste misure di sintesi.

3.20 Il periodico *National Geographic Traveler* ha di recente presentato dei dati riferiti al numero medio annuo di giorni di vacanza trascorsi dai cittadini di otto diverse nazioni: 42 giorni per l'Italia, 37 per la Francia, 35 per la Germania, 34 per il Brasile, 28 per la Gran Bretagna, 26 per il Canada, 25 per il Giappone e 13 per gli USA.

a. Trova la media e la deviazione standard e interpreta i risultati.
b. Riporta la sintesi-a-cinque-numeri (*suggerimento*: puoi calcolare il primo quartile trovando la mediana della prima metà della distribuzione).

3.21 L'indicatore di sviluppo umano HDI (*Human Development Index*) è un indice calcolato dalle Nazioni Unite per formare una graduatoria dei paesi membri sulla base dei valori assunti in ciascuno di essi da grandezze quali la speranza di vita alla nascita, il livello di istruzione e il reddito. Nel 2006 le dieci nazioni con il più alto livello dell'indicatore sono state, in ordine di livello decrescente Norvegia (38), Islanda (33), Australia (28), Irlanda (14), Svezia (45), Canada (24), Giappone (11), USA (15), Svizzera (25), Paesi Bassi (34). Il numero tra parentesi è una misura dell'uguaglianza fra i sessi e indica la percentuale di seggi parlamentari che in ciascun paese è occupata da donne. Calcola la media e la deviazione standard e interpreta i risultati.

3.22 Lo *Human Development Report 2006* pubblicato dalle Nazioni Unite mostra la speranza di vita nelle diverse nazioni. Per l'Europa Occidentale i valori riportati erano: Danimarca 77, Portogallo 77, Paesi Bassi 78, Finlandia 78, Grecia 78, Irlanda 78, Regno Unito 78, Belgio 79, Francia 79, Germania 79, Norvegia 79, Italia 80, Spagna 80, Svezia 80, Svizzera 80.

Tabella 3.14

Vittime	Frequenza	Percentuale
0	1244	90.8
1	81	5.9
2	27	2.0
3	11	0.8
4	4	0.3
5	2	0.1
6	1	0.1

N	Media	Dev. Std.	Max	Q3	Mediana	Q1	Min
1370	0.146	0.546	6	0	0	0	0

Per l'Africa, i valori riportati (molti dei quali sensibilmente più bassi, a causa dell'incidenza dell'AIDS, di quelli registrati cinque anni prima) erano:

Botswana 37, Zambia 37, Zimbabwe 37, Malawi 40, Angola 41, Nigeria 43, Rwanda 44, Uganda 47, Kenya 47, Mali 48, Sud Africa 49, Congo 52, Madagascar 55, Senegal 56, Sudan 56, Ghana 57.

a. Quale gruppo ritieni che abbia la deviazione standard più elevata? Perché?

b. Trova la deviazione standard per ciascun gruppo. Confronta i risultati per mostrare che s è maggiore per il gruppo nel quale si osserva maggiore dispersione.

3.23 Uno studio ha mostrato che nell'Ontario i salari annuali degli insegnanti hanno una media di $ 50 000 e una deviazione standard di $ 10 000 (dollari canadesi). Supponi che la distribuzione sia approssimativamente campanulare.

a. Definisci un intervallo di valori che contiene circa (i) il 68%, (ii) il 95% e (iii) tutte o quasi tutte le osservazioni.

b. Ritieni che un salario pari a $ 100 000 sarebbe anomalo (outlier) nella distribuzione? Perché?

3.24 Escludendo gli USA, la distribuzione del numero medio annuo di giorni di vacanza nei paesi OECD (vedi Problema 3.20) è approssimativamente campanulare con media pari a 35 giorni e deviazione standard pari a 3 giorni[2].

a. Usa la regola empirica per descrivere la variabilità dei dati.

b. Il valore osservato nella distribuzione per gli USA è 13. Se lo includiamo tra le altre osservazioni, (i) la media aumenta o diminuisce, (ii) la deviazione standard aumenta o diminuisce?

c. Utilizzando la media e la deviazione standard calcolata per le altre nazioni determina a quante deviazioni standard dalla media ricade il valore osservato negli USA.

3.25 Nell'ambito della GSS alla domanda "quanta gente conosci che si è uccisa?" l'88% dei rispondenti ha indicato la modalità 0, l'8.8% ha risposto 1 e le restanti risposte hanno indicato valori più elevati. La media della distribuzione è risultata essere pari a 0.145 mentre la deviazione standard a 0.457.

a. Qual è la percentuale di osservazioni che ricade entro una deviazione standard dalla media?

b. La regola empirica è appropriata per questa distribuzione? Perché?

3.26 Un esame è valutato con voti compresi in una scala da 0 a 100; la media è pari a 76. Quale valore è più probabile per la deviazione standard: −20, 0, 10, o 50? Perché?

3.27 I punteggi ottenuti in un esame sono compresi fra 2.0 e 4.0. Considera i seguenti possibili valori per la deviazione standard: −10.0, 0.0, 0.4, 1.5, 6.0.

a. Qual è il valore più realistico per la deviazione standard? Perché?

b. Quale valore è *impossibile*? Perché?

3.28 Secondo quanto riportato dallo US Census Bureau, nel 2005 il prezzo di vendita mediano delle case negli USA era di $ 184 100. Quale dei seguenti valori per la deviazione standard è il più plausibile:
(*a*) −15 000, (*b*) 1000, (*c*) 10 000, (*d*) 60 000, (*e*) 1 000 000? Perché?

3.29 I consumi elettrici residenziali registrati[3] a Gainesville in Florida nel 2006, hanno avuto un valore medio di 10 449 e una deviazione standard di 7489 kilowatt-ora. Il consumo massimo è risultato pari a 336 240 kW/h.

a. Quale ritieni sia la forma della distribuzione? Perché?

b. Ritieni che ci siano outlier nella distribuzione? Spiega.

3.30 I consumi idrici residenziali (misurati in migliaia di galloni[4]) di Gainesville in Florida nel 2006 avevano una media di 78 con una deviazione standard pari a 119. Quale forma ritieni possa avere questa distribuzione? Perché?

3.31 Dallo *Statistical Abstract of the United States 2006* (sommario statistico degli Stati Uniti) risulta che, nel 2004, il salario medio annuo degli insegnanti di scuola superiore negli USA variava tra gli stati con:

100%	Max	61 800 (Illinois)
75%	Q3	48 850
50%	Med	42 700
25%	Q1	39 250
0%	Min	33 100 (South Dakota)

a. Calcola e interpreta il campo di variazione.

b. Calcola e interpreta lo scarto interquartile.

3.32 Facendo riferimento al precedente problema.

a. Traccia il box plot.

b. Sulla base di quanto ottenuto in (*a*), definisci che tipo di asimmetria ha la distribuzione. Spiega.

c. Se la distribuzione, sebbene asimmetrica, fosse grossomodo campanulare quale valore è più plausibile per la deviazione standard: (i) 100, (ii) 1000, (iii) 7000, (iv) 25 000? Spiega.

[2] *Fonte.* Tabella 8.9 in www.stateofworkingamerica.org, The Economic Policy Institute.

[3] Dati forniti da Todd Kamhoot, Gainesville Regional Utilities.

[4] Un gallone USA corrisponde a 3.79 litri; 1000 galloni corrispondono, pertanto, a 3.79 m^3 (*N.d.C.*).

3.33 La Tabella 3.15 mostra parte dell'output ottenuto attraverso un software statistico analizzando i tassi di criminalità (per 100 000 abitanti) riportati nel file di dati "2005 statewide crime" reperibile nel sito web di questo testo. La prima colonna è riferita all'intero dataset mentre la seconda esclude il valore registrato nel District of Columbia (D.C.). Per ciascuna delle statistiche riportate valuta l'effetto esercitato dall'inclusione dell'osservazione outlier.

Tabella 3.15

Variable = MURDER			
N	51	N	50
Mean	5.6	Mean	4.8
Std Dev	6.05	Std Dev	2.57
	Quartiles		Quartiles
100% Max	44	100% Max	13
75% Q3	6	75% Q3	6
50% Med	5	50% Med	5
25% Q1	3	25% Q1	3
0% Min	1	0% Min	1
Range	43	Range	12
Q3-Q1	3	Q3-Q1	3
Mode	3	Mode	3

3.34 Di recente, durante un semestre, nella University of Florida, gli studenti titolari di un'utenza presso il server centrale hanno impiegato una media di 1921 kilobyte di spazio[5]. La deviazione standard è risultata pari a 11 495.

a. Possiamo applicare la regola empirica a questa distribuzione? Perché?

b. La sintesi-a-cinque-numeri è risultata essere: minimo = 4, Q1 = 256, mediana = 530, Q3 = 1105 e massimo = 320 000. Questi valori cosa suggeriscono in merito alla forma della distribuzione?

c. Utilizza il criterio 1.5(IQR) per determinare se nella distribuzione sono presenti osservazioni outlier.

3.35 Per le seguenti distribuzioni ipotizza la forma dell'istogramma e spiega quale tra media e mediana assume i valori più elevati.

a. Il prezzo di vendita delle case di nuova costruzione nel 2008.

b. Il numero complessivo di figli avuti dalle donne di 40 anni e più.

c. Il punteggio in un esame semplice (media = 88, deviazione standard = 10, massimo possibile = 100).

d. Il numero di auto per famiglia.

e. Numero di mesi in cui un soggetto ha guidato un auto nell'ultimo anno.

3.36 Per ciascuna delle seguenti variabili indica se ritieni che il corrispondente istogramma delle frequenze relative sia campanulare, a U, asimmetrico verso destra, asimmetrico verso sinistra.

a. Punteggio di un esame semplice (media = 88, deviazione standard = 10, minimo = 65, primo quartile = 77, mediana = 85, terzo quartile = 91, massimo = 100).

b. Quoziente intellettivo (IQ) per l'intera popolazione.

c. Numero di volte in cui un individuo è stato arrestato nello scorso anno.

d. Tempo necessario per completare un test difficile (tempo massimo 1 ora).

e. Età alla morte.

f. Contribuzione settimanale all'obolo della chiesa di appartenenza (mediana $ 10 e media $ 17).

g. Opinione rispetto alla legalizzazione dell'aborto.

3.37 Per le parti (*a*), (*b*) e (*f*) del problema precedente, rappresenta quello che ritieni possa essere il plausibile box plot della distribuzione.

3.38 Nel gennaio del 2007 il tasso di disoccupazione in 27 paesi dell'Unione Europea assumeva valori compresi fra 3.2 (Danimarca) e 12.6 (Polonia); primo quartile = 5.0, mediana = 6.7, terzo quartile = 7.9, media = 6.7, deviazione standard = 2.2. Rappresenta il box plot etichettando i valori impiegati per costruirlo.

3.39 Con riferimento all'indagine sul numero di quotidiani letti settimanalmente di cui si è parlato nel Problema 1.11, la Figura 3.20 mostra l'output ottenuto attraverso un software statistico. In esso sono rappresentati il diagramma ramo-e-foglie e il box plot.

a. Attraverso il box plot individua il minimo, il primo quartile, la mediana, il terzo quartile e il massimo.

b. Identifica gli stessi cinque valori attraverso il diagramma ramo-e-foglie.

c. Ti sembra che nei dati vi siano outlier?

d. Quale ritieni sia, tra i seguenti − 0.3, 3, 13, 23, il valore della deviazione standard? Perché?

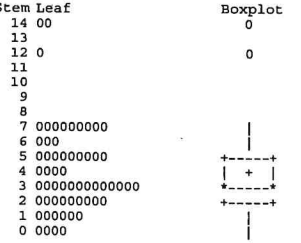

Figura 3.20

3.40 Nel 2006, la distribuzione dei tassi di mortalità infantile (numero di bambini morti entro il primo anno di vita ogni 1000 nati vivi) per i paesi africani pubblicata

[5] Dati forniti da Dr. Michael Conlon, University of Florida.

dalle Nazioni Unite presentava la seguente sintesi-a-cinque-numeri:

$$\text{min} = 54,\ Q1 = 76,\ \text{mediana} = 81$$
$$= Q3 = 101,\ \text{max} = 154$$

Per l'Europa Occidentale la sintesi-a-cinque-numeri era:

$$\text{min} = 3,\ Q1 = 4,\ \text{mediana} = 4,\ Q3 = 4,\ \text{max} = 5.$$

Rappresenta affiancati i box plot per le due distribuzioni e usali per evidenziare le differenze (il diagramma riferito all'Europa mostra che i quartili, al pari della mediana, sono poco utili quando i valori assunti dalle osservazioni sono fortemente discreti).

3.41 Per i diversi stati USA nel 2004 la sintesi-a-cinque-numeri della distribuzione percentuale di individui privi di assicurazione sanitaria aveva un minimo pari all'8.9% (Minnesota), Q1 = 11.6, mediana = 14.2, Q3 = 17.0, massimo 25.0% (Texas) (*Statistical Abstract of the United States, 2006*).

a. Rappresenta il box plot.
b. Ritieni che la distribuzione sia simmetrica, asimmetrica positiva o asimmetrica negativa?

3.42 Nel 2004, negli Stati Uniti, la distribuzione dei voti di diploma di scuola media superiore aveva un minimo pari 78.3 (Texas), primo quartile di 83.6, mediana di 87.2, terzo quartile di 88.8, e massimo pari a 92.3 (Minnesota) (*Statistical Abstract of the United States, 2006*).

a. Determina il campo di variazione e lo scarto interquartile. Interpretali.
b. Si individuano outlier impiegando il criterio del 1.5(IQR)?

3.43 Utilizzando un software analizza i tassi di omicidio riportati nel file di dati "2005 statewide crime" riportato nel sito web del testo.

a. Escludendo dal dataset l'osservazione del District of Columbia (D.C.) trova la sintesi-a-cinque-numeri.
b. Costruisci il box plot e interpretalo.
c. Ripeti l'analisi includendo l'osservazione D.C. e confronta i risultati.

3.44 In un report prodotto dall'OECD[6] veniva indicato che la distribuzione dei consumi idrici annuali dei paesi OECD (vedi Problema 3.6) era asimmetrica positiva; i valori (espressi in metri cubi pro capite) avevano una mediana di circa 500 e un campo di variazione da circa 200 (Danimarca) a 1700 (USA). Considera i seguenti valori per l'IQR: −10, 0, 10, 350, 1500. Qual è il più realistico? Perché?

3.45 Secondo quanto riportato sullo *Human Development Report* pubblicato dalle Nazioni Unite, la distribuzione dei valori delle emissioni di anidride carbonica nel 2005 per le 25 nazioni dell'Unione Europea aveva una media di 8.3 e una deviazione standard di 3.3 tonnellate pro capite. Tutti i valori della distribuzione erano al di sotto di 12 tranne che quello riferito al Lussemburgo che era pari a 21.1.

a. Quante deviazioni standard sopra la media ricade il valore del Lussemburgo?
b. Il valore osservato per la Svezia era 5.8. Quante deviazioni standard sotto la media ricade?
c. Le emissioni di anidride carbonica erano 16.5 per il Canada e 20.1 per gli USA. Con riferimento alla distribuzione delle emissioni nell'Unione Europea trova e interpreta gli z-score per (i) il Canada e (ii) per gli USA.

3.46 Nel volume *Energy Statistics Yearbook* (unstats.un.org/unsd/energy) pubblicato dalle Nazioni Unite, vengono riportati i consumi di energia. Per le 25 nazioni dell'Unione Europea, nel 2006, i valori dei consumi energetici (in kg pro capite[7]) avevano una media di 4998 e una deviazione standard di 1786.

a. L'Italia aveva un valore pari a 4222. A quante deviazioni standard dalla media si trovava?
b. Il valore per gli Stati Uniti era di 11067. Con riferimento alla distribuzione dell'Unione Europea, trova lo z–score per gli USA e interpreta il risultato.
c. Se la distribuzione fosse stata campanulare, il valore di 11067 sarebbe stato anomalo? Perché?

3.47 In uno studio sono state confrontate le opinioni espresse (a favore o contro) da Democratici e da Repubblicani sul servizio nazionale di assicurazione sanitaria.

a. Individua la variabile risposta e la variabile esplicativa.
b. Spiega come i dati potrebbero essere riassunti in una tabella di contingenza.

3.48 La Tabella 3.16 mostra il livello di felicità dichiarato dai rispondenti della GSS che nel 2004 avevano detto di assistere frequentemente o raramente a delle funzioni religiose (le variabili erano ATTEND e HAPPY).

Tabella 3.16

Frequenza funzioni religiose	Molto felice	Abbastanza felice	Non troppo felice	Totale
Ogni settimana o più	200	220	29	449
Mai o meno di una volta l'anno	72	185	53	310

[6] *OECD Key Environmental Indicators 2005.*
[7] Si tratta di un kg equivalente di petrolio.

a. Identifica la variabile esplicativa e la variabile risposta.
b. Per ciascun livello di frequenza alle funzioni religiose trova la percentuale di chi ha detto di essere molto felice.
c. Ti sembra che ci sia un'associazione tra le variabili? Perché?

3.49 Utilizzando dati recenti delle Nazioni Unite riferiti a diversi paesi è stata formulata un'equazione di previsione per la fertilità (numero medio di figli per donna adulta) utilizzando come predittore la percentuale di popolazione che utilizza Internet. L'equazione è:

$$\text{fertilità prevista} = 3.2 - 0.04 \times (\text{uso di Internet})$$

a. Confronta la fertilità prevista in una nazione in cui il 50% della popolazione usa Internet (gli USA) e una nazione in cui lo usa lo 0% (Yemen).
b. La correlazione fra le variabili è −0.55. Spiega cosa rappresenta questo valore negativo.

3.50 Fai riferimento al precedente problema. Un'equazione di previsione in cui la fertilità viene messa in relazione con la percentuale di persone che usa metodi di contraccezione è:

$$\text{fertilità prevista} = 6.6 - 0.065 \times (\text{uso di contraccettivi})$$

mentre la correlazione è −0.89.

a. Che tipo di schema grafico ritieni sia scaturito allo scatterplot ottenuto per questi dati?
b. Quale variabile sembra essere maggiormente associata con la fertilità – Internet o l'uso di contraccettivi? Perché?

3.51 Per i dati delle nazioni OECD riportati nella Tabella 3.11 del Problema 3.6, usa un software statistico per costruire uno scatterplot in cui metti in relazione y = emissioni anidride carbonica e x = PIL.

a. Sulla base di tale scatterplot ti aspetti che la correlazione fra le variabili sia positiva o negativa? Perché?
b. Hai visto qualche osservazione che si discosta particolarmente dalle altre? Identifica la nazione.

3.52 Fai riferimento al precedente problema. La correlazione delle emissioni di anidride carbonica è di 0.03 con il tasso di attività economica femminile e di −0.52 con il numero di medici nella popolazione. Quale variabile è associata più fortemente con le emissioni di anidride carbonica? Perché?

3.53 Qual è la differenza fra le misure descrittive che fanno uso dei simboli

a. $\bar{y}$ e μ? *b.* s e σ?

Applicazioni

3.54 Con riferimento al file di dati "Student survey" reperibile nel sito web del testo (vedi Problema 1.11), utilizza un software statistico per fare delle sintesi grafiche e numeriche delle variabili

a. Distanza dalla città di origine.
b. Ore trascorse guardando la TV in una settimana. Descrivi la forma delle distribuzioni e sintetizza ciò che sei riuscito a evidenziare.

3.55 Fai riferimento al file di dati costruito per il Problema 1.12. Scegli alcune variabili e fai la loro analisi descrittiva. Nel tuo report finale fai un esempio di ipotesi di ricerca che possono essere sviluppate utilizzando le tue analisi; identifica variabili esplicative e variabili risposta. Interpreta i risultati che ottieni.

3.56 La Tabella 3.17 mostra i tassi di decessi (omicidi, suicidi, morti accidentali) per 100 000 abitanti causati da armi da fuoco in diverse nazioni industrializzate. Prepara un report in cui sintetizzi i dati utilizzando i metodi grafici e numerici presentati in questo capitolo.

3.57 Con riferimento al file di dati "2005 statewide crime" reperibile nel sito web del testo, considera il tasso di criminalità e la percentuale di popolazione al di sotto della soglia di povertà. Formula un'ipotesi di ricerca per valutare la direzione della relazione associativa che lega queste variabili identificando la variabile risposta e la variabile esplicativa. Impiegando un software costruisci lo scatterplot e trova la correlazione. Interpreta i risultati e commenta ciò che la correlazione indica in merito alla tua ipotesi di ricerca.

Tabella 3.17

Nazione	Morti per arma da fuoco	Nazione	Morti per arma da fuoco	Nazione	Morti per arma da fuoco
Australia	1.7	Grecia	1.8	Norvegia	2.6
Austria	3.6	Islanda	2.7	Portogallo	2.1
Belgio	3.7	Irlanda	1.5	Spagna	0.7
Canada	3.1	Italia	2.0	Svezia	2.1
Danimarca	1.8	Giappone	0.1	Svizzera	6.2
Finlandia	4.4	Lussemburgo	1.9	Regno Unito	0.3
Francia	4.9	Paesi Bassi	0.8	USA	9.4
Germania	1.5	Nuova Zelanda	2.3		

Fonte. Small Arms Survey, Geneva, 2007.

dono a essere più piccoli impiegando più categorie. Così, in generale, più "raffinata" è la categorizzazione, più stretto è l'intervallo di confidenza per la misura di associazione. Inoltre, misure più raffinate forniscono motivazioni più solide per trattare i dati ordinali come quantitativi e usare, all'evenienza, i metodi più potenti presentati nel capitolo successivo dedicato alle variabili quantitative.

Tabelle di contingenza miste: ordinali-nominali

Per una tabella di contingenza che incrocia una variabile nominale con soltanto due categorie e una variabile ordinale, le misure ordinali di associazione sono ancora valide; in questo caso, il segno della misura indica quale livello della variabile nominale è associato con categorie più elevate della variabile ordinale. Ad esempio, supponiamo che gamma $= -0.12$ per l'associazione in una tabella 2×3 che incrocia sesso (femmine, maschi) e felicità (non troppo felice, abbastanza felice, molto felice). Poiché il segno è negativo, il livello più "elevato" della variabile sesso (cioè, maschio) tende a osservarsi in corrispondenza di più bassi livelli di felicità (l'associazione fra le due variabili è, comunque, piuttosto debole).

Quando la variabile nominale ha più di due categorie, è inappropriato usare una misura ordinale come gamma. Esistono metodi specifici per tabelle di contingenza miste, ovvero con variabili nominali e ordinali, ma è di solito più semplice trattare la variabile ordinale come quantitativa assegnando punteggi ai relativi livelli.

8.7 Riassunto del capitolo

In questo capitolo sono stati presentati alcuni metodi per l'analisi di associazione fra variabili categoriali in tabelle di contingenza.

L'associazione è stata valutata nel modo seguente:

- Attraverso la *descrizione dei conteggi* nella **tabella di contingenza** facendo uso delle distribuzioni percentuali, chiamate **distribuzioni condizionate** percentuali, della variabile risposta. Se le distribuzioni condizionate nella popolazione sono identiche, le due variabili sono **statisticamente indipendenti** – la probabilità di una qualsiasi risposta è la stessa per ciascun livello della variabile esplicativa.
- Attraverso il **chi-quadro** per verificare "H_0: indipendenza" tra variabili. La statistica test χ^2 confronta ogni frequenza osservata f_o alla frequenza attesa f_e sotto H_0, usando

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Per grandi campioni, la statistica test ha la distribuzione campionaria di un chi-quadro. I **gradi di libertà** dipendono dal numero di righe r e dal numero delle colonne c: $gdl = (r-1)(c-1)$. Il P-valore è la probabilità sottesa alla coda destra oltre il valore osservato del test χ^2.

- Attraverso la *descrizione della struttura di associazione* facendo uso dei **residui standardizzati** per le celle della tabella. Un residuo standardizzato rappresenta la distanza, in termini di errore standard, di $(f_o - f_e)$ da 0. Un valore più grande di 2 o 3 in valore assoluto indica che la cella fornisce evidenza di associazione in una particolare direzione.
- Attraverso la *descrizione dell'intensità dell'associazione*. Per tabelle 2×2 la **differenza di proporzioni** è utile, come lo è l'**odds ratio**, il rapporto degli odd di due righe. Ciascun odd è il rapporto fra la proporzione di successi e la proporzione di insuccessi. Quando c'è indipendenza, la differenza di proporzioni è uguale a 0 e l'odds ratio è uguale a 1. Più forte è l'associazione, più lontani sono tali indici dai loro valori di riferimento per l'indipendenza.

Questo capitolo ha anche presentato metodi per analizzare l'associazione tra due variabili ordinali.

- Molte **misure ordinali di associazione** usano il numero delle **coppie concordanti** (il soggetto appartenente a una categoria elevata di x appartiene anche a una categoria elevata di y) e **coppie discordanti** (il soggetto che è classificato in una categoria alta di x è classificato in una categoria bassa di y).
- **Gamma** è uguale alla differenza tra le proporzioni di coppie concordanti e coppie discordanti. Gamma varia tra -1 e $+1$; valori più grandi in valore assoluto indicano associazioni più forti. Quando le variabili sono indipendenti, gamma è uguale a 0.

Il test chi-quadro tratta i dati come nominali. Quando le variabili sono ordinali, i metodi che usano tale ordinamento (come un test z basato sul valore campionario di gamma) sono più potenti per rilevare associazioni caratterizzate da trend positivi o negativi.

Il capitolo successivo presenterà metodi per descrivere e fare inferenza sull'associazione tra due variabili quantitative.

Problemi

Concetti di base

8.1 Le indagini GSS mostrano frequentemente che negli Stati Uniti circa il 40% degli uomini e il 40% delle donne ritengono che una donna possa avere la possibilita di abortire per qualsiasi motivo (variabile ABANY).

a. Costruisci una tabella di contingenza mostrando la distribuzione condizionata della variabile "opinione sull'aborto legale" (sì, no) secondo il sesso.

b. Sulla base di questi risultati, l'indipendenza statistica tra opinione sull'aborto e genere sembra plausibile? Spiega.

8.2 Se una donna sarà incinta o no il prossimo anno è una variabile categoriale con due categorie (sì, no); anche la variabile riferita al fatto se lei o il suo compagno usano contraccettivi è categoriale con due categorie (sì, no). Queste due variabili potrebbero essere statisticamente indipendenti o associate? Spiega?

8.3 Ogni anno, un'indagine in larga scala su matricole universitarie condotta dall'*Higher Education Research Institute* della UCLA (University of California Los Angeles) ha intervistato 283 000 soggetti su molteplici questioni. Nel 2002 il 46% dei maschi e il 35% delle femmine ha indicato di essere favorevole alla legalizzazione della marijuana.

a. Se i risultati a livello di popolazione fossero simili a questi, sesso e opinione sulla legalizzazione della marijuana dovrebbero essere indipendenti o dipendenti?

b. Riporta le percentuali delle popolazioni in una tabella di contingenza sull'ipotesi si indipendenza fra queste variabili.

8.4 Alcuni analisti politici hanno dichiarato che durante la presidenza di George W. Bush la popolarità degli Stati Uniti è molto diminuita nel mondo. In *America Against the World: How We Are Different and Why We Are Disliked*[2], il Pew Research Center ha sintetizzato i risultati di 91 000 interviste condotte in 51 nazioni. In Germania, ad esempio, lo studio ha riportato che la proporzione di coloro che avevano un'opinione

[2] Kohut, A. e Stokes, B. (2006). *America Against the World: How We Are Different and Why We Are Disliked*. Times Books.

positiva degli Stati Uniti è cambiata tra il 2000 e il 2006 dal 78% al 37%. Mostra come costruire una tabella di contingenza che metta in relazione l'opinione sugli USA in Germania con l'anno di intervista. Per tale tabella identifica la variabile risposta, la variabile esplicativa e le distribuzioni condizionate.

8.5 Sulla base delle stime attuali di quanto efficacemente gli apparecchi per mammografia rilevino il cancro alla mammella, la Tabella 8.20 mostra i valori attesi per 100 000 donne adulte quarantenni rispetto alle variabili "Presenza di cancro al seno" e "Risultato della mammografia".

a. Costruisci le distribuzioni condizionate per i risultati del test diagnostico dato il vero stato della malattia. La mammografia risulta essere un buono strumento diagnostico?

b. Costruisci la distribuzione dello stato della malattia per le donne che hanno avuto un risultato del test positivo. Utilizza questo dato per spiegare perché anche un buon test diagnostico può avere un elevato tasso di falsi positivi quando una malattia non è molto diffusa.

Tabella 8.20

		Test diagnostico	
		Positivo	Negativo
Cancro al seno	Sì	860	140
	No	11 800	87 120

8.6 I dati disponibili sul sito web del FBI (www.fbi.gov) indicano che il 91% di tutti i neri assassinati nel 2005, è stato assassinato da neri, mentre l'83% di tutti i bianchi assassinati nel 2005 è stato ucciso da bianchi. Si indichi con y la razza della vittima e con x la razza dell'assassino.

a. A quali distribuzioni condizionate fanno riferimento tali statistiche? Quelle di y a dati livelli di x o quelle di x a dati livelli di y? Costruisci una tabella di contingenza che mostra queste distribuzioni.

b. x e y sono indipendenti o dipendenti? Spiega.

8.7 Quale valore di χ^2 fornisce un P-valore di 0.05 nel verificare l'indipendenza in tabelle delle seguenti dimensioni?

a. 2×2 *b.* 3×3 *c.* 2×5 *d.* 5×5 *e.* 3×9

8.8 Mostra che la Tabella di contingenza 8.21 ha quattro gradi di libertà e indica come i quattro conteggi di cella determinano gli altri.

Tabella 8.21

10	20		60
30	40		100
			40
50	80	70	

8.9 Nel 2000 la GSS ha chiesto se un soggetto è disposto ad accettare tagli del suo standard di vita per aiutare l'ambiente (GRNSOL). Le categorie di risposta erano molto disponibile, abbastanza disponibile, né disponibile né indisponibile, non molto disponibile, per niente disponibile. Quando questa variabile è incrociata con il sesso si ha $\chi^2 = 8.0$.

a. A quali ipotesi si riferisce il test?

b. Riporta i valori di *gdl* su cui χ^2 è basato.

c. A quali conclusioni si giungerebbe utilizzando un livello di significatività di (i) 0.05, (ii) 0.10? Esplicita le tue conclusioni nel contesto di questo studio.

8.10 La Tabella 8.22 si riferisce a una indagine su studenti di una scuola superiore a Dayton, nell'Ohio.

a. Costruisci le distribuzioni condizionate che trattano il fumo di sigaretta come variabile risposta. Interpreta.

b. Verifica se l'abitudine al fumo e all'alcool sono statisticamente indipendenti. Riporta il P-valore e interpreta.

Tabella 8.22

		Uso di sigarette	
		Sì	No
Uso di alcool	Sì	1449	500
	No	46	281

Fonte. Si ringrazia il prof. Harry Khamis per aver fornito questi dati.

8.11 La gente che crede nella vita ultraterrena è più felice? Vai al sito web della GSS sda.berkeley.edu/GSS e scarica la tabella di contingenza, riferita all'indagine del 2006, in cui si mette in relazione felicità e credenza nella vita ultraterrena (variabili HAPPY e POSTLIFE, con YEAR (2006) come filtro di selezione).

a. Definisci una domanda di ricerca che potrebbe essere pertinente con il risultato.

b. Riporta le distribuzioni condizionate usando la felicità come variabile risposta e interpreta.

c. Riporta il valore di χ^2 e il corrispondente P-valore (lo puoi ottenere utilizzando 'Statistics'). Interpreta.

d. Interpreta i residui standardizzati (è possibile ottenerli utilizzando 'z-statistic').

8.12 Nella GSS, soggetti sposati sono stati intervistati sulla felicità del loro matrimonio, variabile codificata come HAPMAR.

a. Vai sul sito sda.berkeley.edu/GSS/ e costruisci una tabella di contingenza per il 2006 mettendo in relazione HAPMAR col reddito familiare misurato come (sopra la media, in media, sotto la media), considerando FINRELA (r: 1-2; 3; 4-5) come variabile di riga e YEAR (2006) nel filtro di selezione. Usa una tabella o un grafico con distribuzioni condizionate per descrivere l'associazione.

b. Attraverso 'Statistics' si ottiene la statistica chi-quadro. Riporta il suo valore, i *gdl*, il P-valore e interpreta.

8.13 Il campione nella Tabella 8.15 è costituito da 157 neri statunitensi. La Tabella 8.23 mostra i conteggi di cella e i residui standardizzati per reddito e felicità per i soggetti bianchi dell'indagine GSS del 2004.

Tabella 8.23

```
Rows: income   Columns: happiness
              not     pretty     very      All
below          62       187        45      294
             5.34      3.43     -7.40
average        47       270       181      498
            -2.73     -0.57      2.53
above          22       127       118      267
            -2.37     -2.88      4.73
All           131       584       131     1059
Cell Contents:       Count
                     Standardized residual
Pearson Chi-Square = 72.15, DF = 4, P-Value = 0.000
```

a. Spiega come interpretare la statistica chi-quadro di Pearson e il corrispondente P-valore.

b. Spiega come interpretare i residui standardizzati nelle quattro celle degli angoli.

8.14 La Tabella 8.24 mostra le analisi effettuate con SPSS per la GSS 2004 per le variabili orientamento politico (party ID) e razza (race).

a. Riporta le frequenze attese per la prima cella e mostra come SPSS le ha ottenute.

b. Verifica l'ipotesi di indipendenza tra l'orientamento politico e la razza. Riporta la statistica test, il P-valore e interpreta.

c. Usa i residui standardizzati (etichettati ADJ RES che sta per 'adjusted residuals') per descrivere la struttura di associazione.

Tabella 8.24

```
            Count     PARTY_ID
            Exp Val
            Adj Res                                  Row
                      democr    indep     repub    Total
RACE      black        250       106        17      373
                     129.1     129.0     114.9
                      14.2      -2.7     -11.9

          white        640       783      1775     2198
                     760.9     760.0     677.1
                     -14.2       2.7      11.9

      Column Total     890       889       792     2571

      Chi-Square     Value      DF       Significance
      ----------     -----      ----     ------------
      Pearson        234.73     2        0.0000
```

8.15 Per una tabella di contingenza 2×4 che incrocia le variabili sesso e religiosità (molto, moderatamente religioso, leggermente e per niente religioso) in una recente GSS, il residuo standardizzato è 3.2 per le donne che sono molto religiose, −3.2 per i maschi che sono molto religiosi, −3.5 per le donne che non sono per niente religiose e 3.5 per i maschi che non sono per niente religiosi. Tutti gli altri residui standardizzati sono compresi tra −1.1 e 1.1. Interpreta.

8.16 La Tabella 8.25 incrocia le variabili felicità (HAPPY) e stato civile (MARITAL) per i dati dalla GSS 2006.

Tabella 8.25

Stato	Molto felice	Abbastanza felice	Non troppo felice
Sposato	600 (13.1)	720 (−5.4)	93 (−10.0)
Vedovo	63 (−2.2)	142 (−0.2)	51 (3.4)
Divorziato	93 (−6.1)	304 (3.2)	88 (3.6)
Separato	19 (−2.7)	51 (−1.2)	31 (5.3)
Mai sposato	144 (−7.4)	459 (4.2)	127 (4.0)

a. Il software fornisce un $\chi^2 = 236.4$. Interpreta.

b. La Tabella 8.25 mostra anche i residui standardizzati in parentesi. Sintetizza l'associazione indicando quali categorie dello stato civile mostrano forte evidenza di (i) più osservazioni e (ii) meno osservazioni nella popolazione nella categoria *molto felice* rispetto a quanto si sarebbe osservato nel caso di indipendenza.

c. Confronta i gruppi di divorziati e sposati attraverso la differenza tra le proporzioni nella categoria *molto felice*.

8.17 In un sondaggio pubblicato su *USA Today* nel luglio 2006, l'82% dei Repubblicani contro il 9% dei Democratici approvava l'operato del presidente George W. Bush. Come può essere caratterizzata l'associazione tra affiliazione politica e l'opinione sull'operato di Bush? Debole o forte? Spiega perché.

8.18 In una recente GSS, la pena di morte per soggetti condannati per omicidio è stata approvata dal 74% dei bianchi e dal 43% dei neri. I favorevoli erano anche il 75% fra i maschi e il 63% fra le femmine. In questo campione, quale variabile è più associata con l'opinione sulla pena di morte? Razza o sesso? Spiega perché.

8.19 Fai riferimento al Problema 8.10, sull'uso di sigarette e di alcool.

a. Descrivi l'intensità dell'associazione usando le differenze nelle proporzioni di fumatori tra chi fa uso e chi fa non uso di alcool. Interpreta.

b. Descrivi l'intensità dell'associazione usando le differenze nelle proporzioni di consumatori di alcool tra fumatori e non-fumatori. Interpreta.

c. Descrivi l'intensità dell'associazione attraverso l'odds ratio. Interpreta. L'odds ratio dipende dalla scelta della variabile risposta?

8.20 La Tabella 8.26 classifica 68 694 passeggeri di automobili e autocarri leggeri coinvolti in incidenti nello stato del Maine classificati a seconda se stavano indossando la cintura e se erano rimasti feriti. Descrivi l'associazione attraverso

a. La differenza tra proporzioni, trattando se 'Ferito' come la variabile di risposta.

b. L'odds ratio.

Tabella 8.26

		Ferito	
		Sì	No
Cintura	Sì	2409	35,383
di sicurezza	No	3865	27,037

Fonte. Si ringrazia la Dr. Cristanna Cook, Medical Care Development, Augusta, Maine, per avere fornito questi dati.

8.21 Il "Substance Abuse and Mental Health Archive", un'indagine nazionale sulla famiglia del 2003 sull'abuso di droga, ha indicato che tra gli statunitensi di età 26-34 anni, il 51% ha usato marijuana almeno una volta nella vita e il 18% cocaina.

a. Trova gli odd di avere usato (i) marijuana, (ii) cocaina. Interpreta.

b. Trova l'odds ratio per confrontare l'uso di marijuana e l'uso di cocaina. Interpreta.

8.22 Secondo il Dipartimento di Giustizia degli USA, nel 2004 il tasso di incarcerazione nelle prigioni della nazione è stato di 1 per 109 residenti maschi, 1 per 1563 residenti femmine, 1694 per 100 000 residenti neri e 252 per 100 000 residenti bianchi (*Fonte.* www.ojp.usdoj.gov/bjs).

a. Trova l'odds ratio tra le variabili incarcerazione e (i) sesso, (ii) razza. Interpreta.

b. Secondo i valori dell'odds ratio, quale delle due variabili ha la maggiore associazione con la variabile incarcerazione, sesso o razza? Spiega.

8.23 Fai riferimento alla Tabella 8.1 sull'orientamento politico e il sesso. Trova e interpreta l'odds ratio per ciascuna sotto-tabella 2 × 2. Spiega perché dalle ultime due colonne si evince che non c'è associazione fra le variabili.

8.24 Nel 2004 la percentuale delle matricole universitarie che concordava con l'affermazione che le relazioni omosessuali dovessero essere proibite per legge era del 38.0% fra i maschi e del 23.4% fra le femmine (www.gseis.ucla.edu/heri/american freshman.html).

a. L'odds ratio è 2.01. Spiega che cosa è sbagliato nella seguente interpretazione: "La probabilità di una risposta sì per i maschi è 2.01 volte la probabilità di un sì per le femmine." Dai la corretta interpretazione.

b. L'odd di un sì è uguale a 0.613 per i maschi. Stima la probabilità di un sì per i maschi.

c. Sulla base del valore dell'odd di 0.613 per i maschi e dell'odds ratio di 2.01, mostra come stimare la probabilità di un sì per le donne.

8.25 La Tabella 8.27 incrocia le variabili felicità e reddito familiare per il sub-campione di individui, intervistati nella GSS 2004, che si sono dichiarati ebrei.

a. Trova il numero di (i) coppie concordanti, (ii) coppie discordanti.

b. Calcola gamma e interpreta.

c. Mostra come esprimere gamma quale differenza tra due proporzioni.

Tabella 8.27

		HAPPY		
		Not_too	Pretty	Very
INCOME	Below	1	2	1
	Average	0	5	2
	Above	2	4	0

8.26 Per i dati della GSS 2006, risulta $\hat{\gamma} = 0.22$ con riferimento alla relazione tra soddisfazione sul lavoro (SATJOB; con categorie molto insoddisfatto, poco insoddisfatto, moderatamente soddisfatto, molto soddisfatto) e reddito familiare (FINRELA; sotto la media, in media, sopra la media).

a. Come potrebbe essere considerata tale associazione, molto forte o relativamente debole? Spiega.

b. Delle coppie che sono concordanti o discordanti, quale è la proporzione di concordanti? e di discordanti?

c. L'associazione risultante è più forte o più debole di quella tra soddisfazione sul lavoro e felicità (variabile HAPPY), che ha $\hat{\gamma} = 0.40$? Spiega.

8.27 Uno studio sulle aspirazioni riferite al titolo di studio degli studenti di scuola media superiore[3] ha misurato le aspirazioni attraverso la scala (scuola superiore non terminata, scuola superiore terminata, università non terminata, università terminata) e reddito familiare con tre categorie ordinali. Un software ha fornito i risultati mostrati nella Tabella 8.28.

a. Usa gamma per sintetizzare l'associazione.

b. Verifica l'indipendenza tra aspirazioni e reddito familiare attraverso il test del chi-quadro. Interpreta.

c. Trova l'intervallo di confidenza al 95% per gamma. Interpreta.

d. Conduci un test alternativo di indipendenza che prenda in considerazione l'ordinamento tra categorie. Perché i risultati sono così diversi da quelli ottenuti attraverso il test del chi-quadro?

Tabella 8.28

Statistic	DF	Value	Prob
Chi-Square	6	8.871	0.181

Statistic	Value	ASE
Gamma	0.163	0.080

8.28 In riferimento al Problema 8.13 sulla felicità e il reddito, le analisi fatte non prendevano in considerazione la natura ordinale delle variabili. Attraverso l'impiego di un software

a. discuti sull'intensità dell'associazione ricavando e interpretando gamma,

b. costruisci e interpreta un intervallo di confidenza al 95% per il valore di gamma nella popolazione.

Applicazioni

8.29 Facendo riferimento al file di dati "Student survey" (Problema 1.11), costruisci e analizza in modo descrittivo e inferenziale la tabella di contingenza che mette in relazione l'opinione sull'aborto e (*a*) l'affiliazione politica, (*b*) la religiosità.

8.30 Facendo riferimento al file di dati costruito nel Problema 1.12, utilizzando le variabili scelte dal tuo docente definisci un problema di ricerca, conduci un'analisi statistica sia descrittiva sia inferenziale; interpreta e sintetizza i risultati in un breve rapporto.

8.31 La GSS del 2002 ha chiesto agli intervistati come il lavoro domestico venga ripartito tra il rispondente e il corrispondente coniuge (HHWKFAIR). Possibili risposte sono 1 = Io faccio molto di più della parte che mi spetterebbe, 2 = Io faccio un po' di più della parte che mi spetterebbe, 3 = Io faccio la parte che mi spetta, 4 = Io faccio un po' meno della parte che mi spetterebbe, 5 = Io faccio molto meno della parte che mi spetterebbe. La Tabella 8.29 mostra i risultati secondo il sesso del rispondente. Definisci una domanda di possibile interesse di ricerca che potrebbe essere discussa con questo output e scrivi una relazione di una pagina per sintetizzare che cosa è stato appreso ("*Adj. Residual*" indica il residuo standardizzato).

8.32 Stabilisci un quesito di ricerca riguardante l'opinione verso le relazioni omosessuali e l'ideologia politica. Utilizzando i dati dalla GSS più recente su HOMOSEX e POLVIEWS, conduci un'analisi descrittiva e inferenziale per rispondere al quesito e scrivi una breve relazione per sintetizzare le analisi effettuate.

Tabella 8.29

		HHWKFAIR					Total
		1	2	3	4	5	
sex female	Count	121	108	135	19	6	389
	% within sex	31.1%	27.8%	34.7%	4.9%	1.5%	100.0%
	Adj. Residual	8.0	5.9	-4.2	-7.1	-4.9	
male	Count	18	28	148	68	29	291
	% within sex	6.2%	9.6%	50.9%	23.4%	10.0%	100.0%
	Adj. Residual	-8.0	-5.9	4.2	7.1	4.9	

	Value	df	Asymp. Sig.
Pearson Chi-Square	155.8	4	.000

	Value	Asymp. Std. Error
Gamma	.690	.038

[3] S. Crysdale, *Intern. J. Compar. Sociol.*, vol. 16, 1975, pp. 19-36.